This document is published in:

Ferrández Vicente, J. M. et al. (eds.), (2013). *Natural and Artificial Computation in Engineering and Medical Applications: 5th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2013, Mallorca, Spain, June 10-14. Proceedings, Part II*. (Lecture Notes in Computer Science: 7931) Springer, 149-158.
DOI:http://www.dx.doi.org/10.1007/978-3-642-38622-0_16

A Data Fusion Perspective on Human Motion Analysis Including Multiple Camera Applications

Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and José M. Molina

Computer Science Department. Universidad Carlos III de Madrid
Avda. de la Universidad Carlos III, 22
28270 Colmenarejo (Madrid). Spain
{rcilla,mpatrici}@inf.uc3m.es, {aberlan,molina}@ia.uc3m.es

Abstract. Human motion analysis methods have received increasing attention during the last two decades. In parallel, data fusion technologies have emerged as a powerful tool for the estimation of properties of objects in the real world. This papers presents a view of human motion analysis from the viewpoint of data fusion. JDL process model and Dasarathy's input-output hierarchy are employed to categorize the works in the area. A survey of the literature in human motion analysis from multiple cameras is included. Future research directions in the area are identified after this review.

Keywords: Human Action Recognition, Data Fusion, Computer Vision.

1 Introduction

The recognition of human movements [1] has been studied by the computer vision community for more than twenty years. The developments made during this period have enabled the creation of multiple systems. Automatic Surveillance [2], Ambient Intelligence [8] or Human Computer Interaction [5] are some of them. Abnormal behavior detection is employed in Video Surveillance Systems to detect suspicious behaviors that might be assessed as a thread. Smart home environments analyze actions and mood of the inhabitants to adapt the environment to their preferences, changing music or lighting conditions to make it more comfortable. Commercial gaming platforms employ advanced sensors to capture the real movements of the players, providing an enhanced and more realistic experience.

The aim of human movement analysis systems is to transform the pixel intensities in the input video sequences into a semantic intepretation of them. The interpretation might be defined at different knowledge levels. Aggarwal and Cai [1] propose a hierarchy of *gestures*, *actions*, *interactions* and *group activities*. *Gestures* are dened as elementary movements of a persons body part, and are the atomic components describing the meaningful motion of a person. *Actions* are dened as single-person activities that may be composed of multiple gestures

organized temporally. *Interactions* are human activities that involve two or more persons and/or objects. *Group activities* are dened as the activities performed by conceptual groups composed of multiple persons and/or objects. These levels should not be interpreted as closed sets, as many times it is not clear at what level operates a given system.

The first human motion analysis systems developed where limited to the usage of a single camera view. However, in recent years, with the aim of deploying human movement analysis systems in the real world, human movement analysis systems have incorporated multiple camera views, as they provide different advantages:

- Viewpoint invariance. The appearance of actions changes according to the orientation in the execution in the action with respect to the camera. Thus, employing multiple views provides complementary information to achieve a more robust recognition.
- Robustness towards occlusions. In real environments there is usually multiple furniture, walls or other objects that produce partial occlusions in the observed target. The way to overcome this limitation and not loss important motion information is to observe the scene from multiple viewpoints.
- Wider scene coverage. A single camera has a very limited coverage. Multiple cameras are needed to cover full scenes.

Data Fusion studies the efficient combination of measurements obtained from multiple sensors or, alternatively, the temporal measurements obtained from a single sensor, in order to achieve more specific inferences about the state of one or more entities than the ones that could be achieved by using a single, independent, sensor [14]. Human movement analysis systems are covered by this definition, independently of the number of cameras employed and the level of abstraction where the analysis is made. However, to the best of our knowledge, the recognition of human movement has not been studied from the viewpoint of data fusion. The purpose of this paper is to analyze human movement analysis applications from the viewpoint of data fusion.

1.1 Contributions

The contributions of this paper might be summarized as:

- A review of relevant data fusion concepts and frameworks.
- A characterization of Human Action Recognition systems from the viewpoint of the JDL process model.
- A survey of the literature of human action recognition from multiple cameras employing the taxonomy provided by Dasarathy's input-output framework.

1.2 Paper Organization

Paper is organized as follows. Section 2 presents the main concepts and frameworks developed by the data fusion community. Section 3 studies the relationship

between the different data fusion levels and human action recognition. Section 4 surveys the area of human action recognition from multiple cameras. Section 5 concludes the paper discusing about hypothetical ways of colaboration between data fusion and human action recognition communities

2 Data Fusion

The Joint Directors of Laboratories Data Fusion Working Group currently defines Data Fusion as *The process of combining data or information to estimate or predict entity states* [23]. This definition is generic enough to cover a wide range of data association and combination problems appearing on different domains. Data fusion is not a discipline by itself, nor the combination of signal processing, artificial intelligence, estatistica estimation or systems enginering to solve state estimation problems.

Different frameworks have been developed to categorize data fusion systems: the JDL process model and Dasarathy's input-output model. These are complementary frameworks for the analysis of data fusion systems whose usage is widely extended. Both are introduced in next paragraphs an will be later employed for the analysis of human movement analysis systems.

2.1 The JDL Process Model

The JDL Data Fusion model [25] is the most widely used framework for the categorization of data fusion systems and algorithms. The first version was published in 1985 by the US Joint Directors of Laboratories (JDL) Data Fusion Working Group with the aim of providing a common framework to facilitate the communication between the communication between data fusion stakeholders and provide a conceptual framework for new developments. The JDL model is not an architectural paradigm nor a process model for the creation of data fusion system. Instead, it provides different levels of abstraction where the different algorithms employed in data fusion systems might be accommodated according to the kind of processing they perform.

The JDL data fusion model, after the 1998 revision [23], proposes five different levels of abstraction where the data fusion functions are accommodated (figure 1. These levels are:

- Level 0. *Signal/Feature Assessment.* This level includes the algorithms employed to enhance or combine the input signals of the fusion systems. The inferences made at this level do not make any assumption about the causes originating the signals. Typical operations at this level include spatial and temporal data alignment, data standardization and data preconditioning for bias removal.
- Level 1. *Entity Assessment.* Algorithms employed for the estimation of the current state of a individual entities are defined at this level. This includes target detection, classification, location, tracking and identity estimation.

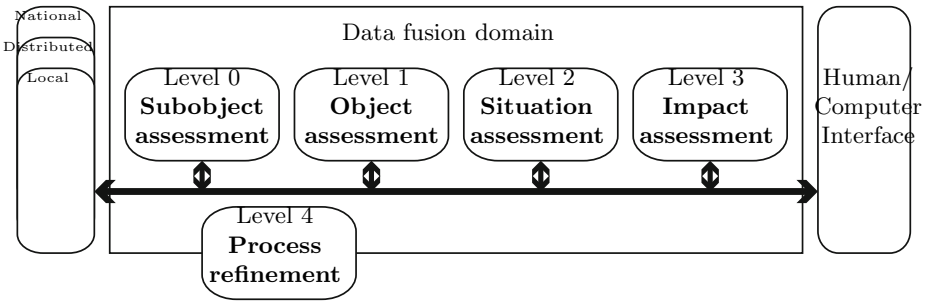

Fig. 1. The JDL data fusion model (1998 revision)

Processing at this level usually implies the association of observations to the corresponding responsible targets.

- Level 2. *Situation Assessment*. A situation is *a set of entities, their attributes, and relationships*. Thus, the task at this processing level is to infer the existent relationships between the analyzed entities employing the individual state estimations.
- Level 3. *Impact Assessment*. The purpose of the algorithms defined at this level is to predict future situations derived from the current and past inferred situations. This includes the computation of expected outcomes of actions executed to alter the current situation or the projection of the current situation to the future to predict the possible evolution.
- Level 4. *Process Assessment*. This level includes the algorithms employed to measure the real-time performance of the fusion system and improve it. This includes the reconfiguration of the sensors employed or the replacement of data fusion algorithms by others better adapted to the current or expected scenario.

2.2 Dasarathy's Input-Output Model

Dasarathy proposed an alternative categorization of Data Fusion systems according to the level of abstraction of the information at the input and output of the fusion system [6]. Three different levels of abstraction are defined: (1) *data*; (2) *features* and (3) *decisions*. Data is the lowest level of abstraction, corresponding to the raw measurements of the sensors, such pixel intensities or depth information. Features are transformations of the data to enhance some property such edges or curvature. Finally, decisions encode information about the certainty of a fact, in the form, among others, of probability estimates or fuzzy sets.

Data fusion systems are characterized according to this abstraction of their inputs and outputs as follows (figure 2):

- Data in-Data out (DAI-DAO) Fusion. At the lowest level of abstraction are systems processing *data* and generating *data*. An example of this kind of fusion systems are multispectral imaging devices: pixel intensities are captured

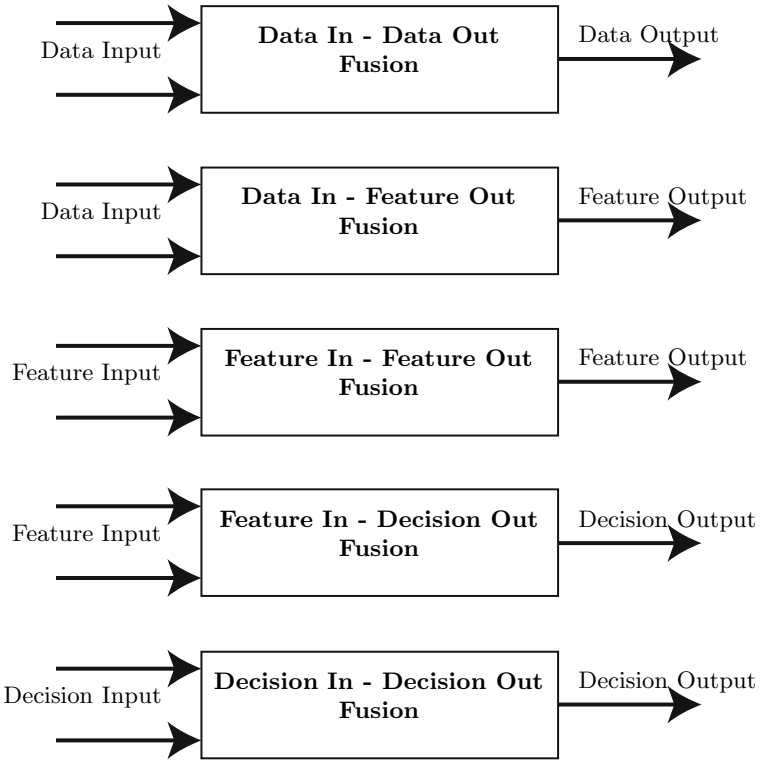

Fig. 2. Dasarathy Input-Output model

at different wavelengths to compose an image better describing the reality. High Dynamic Range (HDR) imaging is another example of a DAI-DAO fusion system, combining images taken with different exposition configurations to have a better representation of the details of dark and light regions of the scene.

- Data in- Feature Out (DAI-FEO) Fusion. At the next level of abstraction in the hierarchy are the systems processing *data* to generate *ig*features. Stereo vision systems are located at this level, as they compute disparity maps (*features*) from pixel intensities (*data*).
- Feature in-Feature Out (FEI-FEO) Fusion. At the mid level of the hierarchy are located systems processing *features*. The conceptually simpler are those generating *features* too. Due to the vague definition of what is a feature at this category lie a wide variety of systems. Fusion systems combining the measurements of the same state variable to provide a more robust estimation of the real value belong to this category.
- Feature In-Decision Out (FEI-DEO) fusion. The next abstraction level is related to pattern recognition systems, transforming *features* into *decisions*

about the class of the phenomena being recognized. At this level are defined those data fusion systems based on introducing a set of features computed from multiple sources into a classifier.
- Decision In-Decision Out (DEI-DEO) fusion. The highest level of abstraction includes the system that combine independent decisions about the phenomena to study to make a global decision about it. Decisions might be defined in different forms, such crisp values, probabilistic distributions or fuzzy sets.

3 Human Action Recognition and the JDL Process Model

This section analyzes human action recognition applications from the view point of the JDL process model. Next paragraphs analyze the relationship of JDL with different human movement analysis applications. JDL levels are confronted with the different abstractions presented at the introduction.

At JDL level 0 are image and video processing methods auxiliarly employed to enhance specific properties of input video sequences, but its definition does not allow to include any specificic method for human motion analysis.

Human movement analysis algorithms analyzing gestures and actions are defined at JDL level 1. The state variable to infer is a label characterizing the kind of action or gesture. This level contains most of the works defined for human motion analysis, as *gesture* and *action* are the better studied abstraction levels. JDL level 1 also includes *group activities*, as the group performing the movement is considered as a whole.

The recognition of *interactions* is performed at JDL level 2. Interactions might be human-human or human-object.

Level 3 in Human Action Recognition corresponds to the prediction of the future actions that person is going to do. However, to the best of our knowledge no applications at this level have been defined. The plan recognition problem [13], where the objective is to infer what is goal of an observed agent would be the closer sample to this level.

Levels 4 and 5 of the JDL process models have not been very exploited from human the human action recognition perspective. Level 4 studies how the information is presented to the system operator. Commercial video surveillance applications incorporate this capabilities, incorporating semantic information in the reports. Commercial gaming platforms with visual inputs represent the motions performed by the player with avatars. Fitness trainers represent with them how the player is performing a given exercise and how they should do it, in order to correct their performance and prevent hurts.

Level 5 would study the adaption of the algorithms employed to new conditions of the environment, such lighting or occlusions. However, to the best of our knowledge, no works have been reported proposing such applications.

4 Human Action Recognition from Multiple Cameras and Dasarathy's Input-Output Model

Dasarathy's input-output model introduced in section 3 provides a framework to categorize the works in Human Action Recognition employing multiple views of the scene being analyzed.

Human Action Recognition methods employing multiple cameras are defined at FEI-FEO, FEI-DEO and DEI-DEO levels. Although fusion at the data levels might be employed for human action recognition, they are not considered, as this kind of fusion is independent of the higher level task.

Diverse methods have been defined at the FEI-FEO data fusion level to combine the information obtained from multiple cameras. Different strategies have been defined at this level. It is possible to divide this works in three different categories: (1) methods projecting 2D features to 3D; (2) methods combining features in a subspace; (3) methods selecting the best available view.

Different 3D representation might be obtained from projecting 2D features to 3D. A popular approach is to recover the 3D shape projecting 2D silhouettes and recovering the visual hull[7,18,17]. Visual hull reconstruction requires accurate silhouette segmentation at the different available views. Recent works have proposed alternatives based on the projection of optical flow to 3D [9], or the projection of local interest points [10]. Other works recover the 3D star skeleton by the correspondence of the corresponding 2D skeletons [3]. The correspondence between action sketches might be computed from multiple views [27]. The main drawback of all these approaches is that they need from accurate camera calibration parameters to perform the projection of the features in 3D.

Alternative methods compute features for the 2D views available and combine them employing some simple scheme. The averaging of the multiple features representing pose, global and local motion has been proposed improving the results with respect to other alternatives [15]. A joint Bag-of-Words histogram might be constructed with the local feature descriptors obtained for each one of the views [26], but a higher performance is obtained with other fusion strategies. Projections maximizing the cross-covariance between the $\mathcal{R}$-transform derivatives computed at each view have been defined to learn a joint subspace where the action recognition is performed [12]. Two level Linear Discriminant Analysis is employed to learn silhouette projections maximizing the separability of the action classes [11]. Cilla et al. proposed variations of Canonical Correlation Analysis to perform the fusion of the diferent motion descriptors computed from the different views [4]. All this methods provide more flexible solutions for the combination of the features obtained from multiple cameras. However, the experimental results show a lower performance than the methods based on 3D reconstruction.

The last class of methods is based on computing a measurement of the quality of each view available, in order to select the best and perform the recognition with the data from that view. A first approach to the selection of the best view is made estimating the orientation of the human with respect to the camera [21]. A measurement based on the properties of the silhouette has been proposed [15].

Other proposed measure in the case of employing local features is to choose the camera with the highest number of detections [26]. Different utility measures have been proposed studying the saliency, concavity or variations of silhouette stacks [20]. The main drawback of this approaches is that they do not exploit the complementary information that might be present at each view.

The next category of works examined employing multiple views of the scene for the recognition of human actions are those defined at the FEI-DEO level. This works model the existing correlations among the multiple observations in the structure of the classifier employed for the prediction of the actions. The concatenation of the input features is the most straightforward procedure to perform the fusion [26,15]. The Fused HMM [24] proposes to model correlations among observations coupling the values of the hidden state chains of parallel HMMs defined for each view. Histograms of local features have been fused rotating the ordering of the inputs to account for the variations in the orientation of the inputs [22]. The main drawback of this works is their lack of flexibility, assuming that the camera configurations remain unchanged between train and test steps. A procedure for the alignment of camera views where the configuration changes from train to test steps is defined in [19], but requiring the knowledge of relative camera placement.

The last category of works employing multiple views performs the fusion at the DEI-DEO level, combining the outputs of action classifiers applied to each one of the camera views. Majority voting has been the most common technique for the fusion of decisions [15,16]. A weighted voting strategy has been proposed in [28], correcting each vote according to the value of the observed feature. Cilla et al. [5] have proposed to learn an error model to weight the predictions made from the different cameras, improving the overall result.

5 Conclusions

This work has analyzed human movement understanding applications employing data fusion concepts and frameworks. The different levels of the JDL process model have been compared to the different steps needed to perform human action recognition. It has been shown that most of the human action recognition algorithms are defined at JDL level 1. At level 2 are defined algorithms studying interactions. Other levels have not been really exploited and they should be targets of future research.

Dasarathy's Input-Ouput hierarchy has been employed to categorize multi-camera human action recognition applications. Existing works have been categorized under three conceptual classes according to the data abstractions employed.

It is clear from this work the existing relationships between data fusion and human movement analysis. However, human movement analysis applications have not been developed according to data fusion practices. Future works will have to exploit these potential sinergies to improve human movement analysis systems.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis. ACM Computing Surveys 43(3), 1–43 (2011)
2. Castanedo, F., Gomez-Romero, J., Patricio, M.A., Garcia, J., Molina, J.M.: Distributed data and information fusion in visual sensor networks. In: Distributed Data Fusion for Network-Centric Operations, p. 435 (2012)
3. Chen, D., Chou, P.C., Fookes, C.B.: Multi-view human pose estimation using modified five-point skeleton model, pp. 17–19 (2008)
4. Cilla, R., Patricio, M.A., Berlanga, A., Molina, J.M.: Multicamera action recognition with canonical correlation analysis and discriminative sequence classification. In: Ferrández, J.M., Álvarez Sánchez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2011, Part I. LNCS, vol. 6686, pp. 491–500. Springer, Heidelberg (2011)
5. Cilla, R., Patricio, M.A., Berlanga, A., Molina, J.M.: A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views. Neurocomputing 75(1), 78–87 (2012)
6. Dasarathy, B.V.: Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proceedings of the IEEE 85(1), 24–38 (1997)
7. Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: The i3DPost Multi-View and 3D Human Action/Interaction Database. In: 2009 Conference for Visual Media Production, pp. 159–168 (November 2009)
8. Gómez-Romero, J., Serrano, M.A., Patricio, M.A., García, J., Molina, J.M.: Context-based scene recognition from visual data in smart homes: an information fusion approach. Personal and Ubiquitous Computing, 1–23 (2011)
9. Holte, M.B., Chakraborty, B.: A Local 3D Motion Descriptor for Multi-View Human Action Recognition from 4D Spatio-Temporal Interest Points, vol. (c), pp. 1–13 (2011)
10. Holte, M.B., Moeslund, T.B., Nikolaidis, N., Pitas, I.: 3D Human Action Recognition for Multi-view Camera Systems. In: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, pp. 342–349 (May 2011)
11. Iosifidis, A., Tefas, A., Nikolaidis, N., Pitas, I.: Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis. Computer Vision and Image Understanding 116(3), 347–360 (2012)
12. Karthikeyan, S., Gaur, U., Manjunath, B.S., Grafton, S.: Probabilistic subspace-based learning of shape dynamics modes for multi-view action recognition. In: 2011 IEEE International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1282–1286 (November 2011)
13. Kautz, H., Allen, J.F.: Generalized plan recognition. In: Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia, PA, vol. 19, p. 86 (1986)
14. Liggins, M.E., Hall, D.L., Llinas, J.: Handbook of multisensor data fusion: theory and practice, vol. 22. CRC (2008)
15. Määttä, T., Aghajan, H.: On efficient use of multi-view data for activity recognition, pp. 158–165 (2010)
16. Naiel, M.A., Abdelwahab, M.M.: Multi-view Human Action Recognition System Employing 2DPCA Motaz El-Saban, pp. 270–275 (2010)
17. Pehlivan, S., Duygulu, P.: A new pose-based representation for recognizing actions from multiple cameras. Computer Vision and Image Understanding 115(2), 140–151 (2011)

18. Peng, B., Qian, G., Rajko, S.: View-invariant full-body gesture recognition via multilinear analysis of voxel data. In: 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC, pp. 1–8 (August 2009)
19. Ramagiri, S., Kavi, R., Kulathumani, V.: Real-time multi-view human action recognition using a wireless camera network. In: 2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–6 (August 2011)
20. Rudoy, D., Zelnik-Manor, L.: Viewpoint Selection for Human Actions. International Journal of Computer Vision 97(3), 243–254 (2011)
21. Shen, C., Zhang, C., Fels, S.: A Multi-Camera Surveillance System that Estimates Quality-of-View Measurement. In: 2007 IEEE International Conference on Image Processing, pp. III-193–III-196 (2007)
22. Srivastava, G., Iwaki, H., Park, J., Kak, A.C.: Distributed and lightweight multi-camera human activity classification. In: 2009 Third ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC, pp. 1–8 (August 2009)
23. Steinberg, A.N., Bowman, C.L., White, F.E.: Revisions to the JDL data fusion model. American Inst. of Aeronautics and Astronautics, New York (1998)
24. Wang, Y., Huang, K., Tan, T.: Multi-view Gymnastic Activity Recognition with Fused HMM, pp. 667–677 (2007)
25. White, F., et al.: A model for data fusion. In: Proc. 1st National Symposium on Sensor Fusion, vol. 2, pp. 149–158 (1988)
26. Wu, C., Khalili, A.H., Aghajan, H.: Multiview activity recognition in smart homes with spatio-temporal features. In: Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2010, p. 142 (2010)
27. Yan, P., Khan, S.M., Shah, M.: Learning 4D action feature models for arbitrary view action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (June 2008)
28. Zhu, F., Shao, L., Lin, M.: Multi-View Action Recognition Using Local Similarity Random Forests and Sensor Fusion. Pattern Recognition Letters (May 2012)